Null-model Validation of MLP Input Contribution Analysis in Ecology

Michael J. Watts and S.P. Worner
National Centre for Advanced Bio-Protection Technologies
PO Box 84
Lincoln University
Lincoln 7647
New Zealand

Abstract

A method is presented for applying a null-model analysis to the verification of the significance of the input neurons of Multi-Layer Perceptrons (MLP). This method was applied to a problem from ecology, namely the establishment of invasive insect pest species. Previous work has described how the MLP were trained to predict species establishment from climate data, and to identify which climatic factors are significant. The null-model analysis method described here was used to validate these predictions.

1 Introduction

There are several advantages to the identification of the contributions of input neurons of an MLP. By identifying inputs that are not significant to the problem, these variables can be removed from the model, yielding a network that is more parsimonious. Conversely, the identification of variables that are significant to the problem is of great value in data mining.

Many methods have been proposed to determine the importance of each of the input neurons of an MLP. These include the methods of Garson [2], Milne [4] and Olden and Jackson [5].

Input contribution analysis is particularly useful when data mining large data sets using a large number of variables, such as those that may be considered important in ecological modelling problems. This paper describes a null-model based method for validating the importance of variables identified by contribution analysis. This technique can be used to validate the results of any input contribution analysis method, in any application area.

A previous study [6] used MLP and contribution analysis to determine which abiotic factors appeared to explain the establishment of several insect pest species in different geographical regions. This paper describes how the importance of these variables was validated by null-model analysis.

2 Method

Data consisting of 135 variables describing the climate in 459 geographical regions as well as the presence and absence of certain insect pest species within those regions were prepared. The pest presence and absence data was extracted from [1]. Conventional statistical modelling of this data had previously proved unsuccessful. In this study twenty percent of the data was held out as an independent validation set, while the remaining eighty percent was used to construct MLP models that predicted the presence or absence of four insect pest species within these regions. These species are listed in Table 2. *M. persicae* and *B. brassicae* are recorded as being present in New Zealand, while *S. zeamais* and *D. melanogaster* are not. For each species, one thousand training and testing sets were randomly drawn from the eighty percent portion of the data set, in a two to one ratio, and used to train and test a three-neuron-layer MLP, with the goal of constructing a model that could predict the presence and absence of the target species with reasonable precision. The MLP with the best performance over the testing set was then evaluated over the validation data set.

The contribution of each input neuron was calculated at the termination of training each network, using the method of Olden and Jackson [5], and the inputs ranked according

Table 1. Target species

Species	Common name
Myzus persicae	Green Peach Aphid
Brevicoryne brassicae	Cabbage Aphid
Sitophilus zeamais	Greater Grain Weevil
Drosophila melanogaster	Common Fruit Fly

Proceedings of the Sixth International Conference on Hybrid Intelligent Systems (HIS'06)
0-7695-2662-4/06 $20.00 © 2006 IEEE

to their mean contributions. The results of this ranking and the performance of the resulting networks are reported in [6].

To validate these analyses, a statistically-solid, null-model approach was developed. Null-models are commonly used in the analysis of ecological data, especially in the analysis of species assemblages [3]. In these analyses, a matrix of species presence and absence is randomly rearranged and compared to the original matrix. If there are significant differences between a sufficient number of random (or null) matrices and the original observed matrix, then the species assemblages are said to not be random. This principle can be applied to the validation of the input contributions of MLP.

Firstly, MLP were constructed and trained on sub-sets of inputs selected according to the mean contributions of the inputs. The inputs were selected so that the sum of their absolute mean contributions was equivalent to specified percentages of the total absolute contribution of all inputs. In other words, the sum of the absolute contribution of all inputs was first calculated, and the inputs placed in descending order of mean absolute contribution. Inputs were then selected one at a time, until the sum of the contributions of the selected inputs equalled or exceeded the selected percentage of the sum of all contributions. Thus, the number of inputs that were selected varied for each species. The percentages of the total input contributions that were selected were $10, 20, 30 \ldots 90\%$ (Table 2).

Secondly, MLP were constructed and trained on an identical number of randomly selected input variables. This comprised the null model of the analysis. The null hypothesis was that there was no significant difference between the performances of MLP trained on randomly selected subsets of inputs, and the performances of MLP trained on subsets of inputs that were selected according to the results of the input contribution analysis. In other words, if there was no significant difference between the performance of the MLP trained using the selected inputs and the MLP using randomly selected inputs, then the method of selecting inputs could be said to be no better than random.

Note that while this method was applied to a problem from ecology, there is no reason why it could not be applied to problems in other disciplines.

3 Results

The mean and standard errors of the accuracies over the training data, as measured by Cohen's Kappa statistic, for each set of features is plotted in Figure 1 for *B. brassicae*, Figure 2 for *M. persicae*, Figure 3 for *S. zeamais* and Figure 4 for *D. melanogaster*. In each of these plots, the mean accuracy of the networks trained using the full variable set is superimposed as a horizontal line. Training accuracy, as opposed to generalisation accuracy, is presented because the goal of these experiments was to investigate how the input variables effected the learning of the network.

Table 2. Number of variables selected by percent contribution and results of hypothesis tests.

B. brassicae									
%	10	20	30	40	50	60	70	80	90
#	5	10	17	25	35	45	57	72	92
H_0	r	r	r	r	r	r	r	r	r
M. persiace									
%	10	20	30	40	50	60	70	80	90
#	4	9	16	24	33	43	54	68	86
H_0	r	r	r	r	r	r	r	r	r
S. zeamais									
%	10	20	30	40	50	60	70	80	90
#	5	10	17	25	34	43	54	66	85
H_0	r	r	r	r	r	r	r	r	r
D. melanogaster									
%	10	20	30	40	50	60	70	80	90
#	4	9	16	23	31	41	52	68	87
H_0	r	r	r	r	r	r	r	r	r

The number of input features selected for each percent contribution is listed in Table 2 for each species, where the rows captioned "%" are the percent contributions and the rows captioned "#" are the number of input variables selected. The rows captioned "H_0" present the results of a statistical test (two tailed t-test, $p = 0.001$) for equivalence between the accuracies of the networks trained using variables selected by their contribution and the accuracies of networks trained using randomly selected variables. In this row, "r" indicates that the null hypothesis was rejected (that is, there was a significant difference between the two sets of accuracies).

Inspection of Table 2 showed that for all species, the accuracies were significantly different for all sets of variables. Inspecting the plots in Figures 1-4 showed that in each case, the accuracy of the networks trained using variables selected by their contributions was greater than that of networks trained using variables selected randomly. These results strongly reject the null hypothesis that the selection of variables using the contribution analysis is no better than the random selection of variables.

To verify that the results of the random feature sets were not distorted by biased random sampling, additional tests were performed by calculating the degree of similarity between the randomly selected features and the features selected by contribution. These similarities were then com-

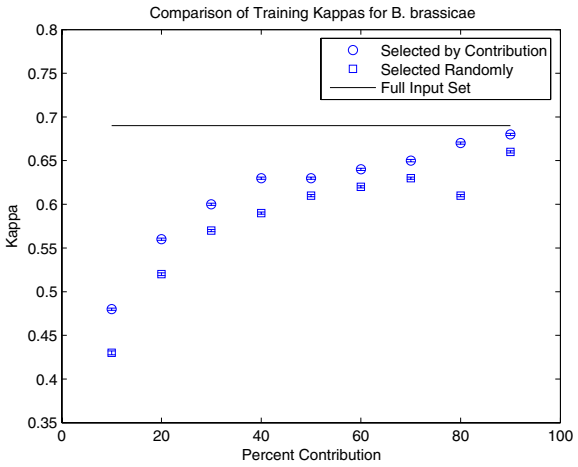

Figure 1. Mean and standard errors of training set kappas for input subsets selected by contribution and selected randomly for species *B. brassicae*

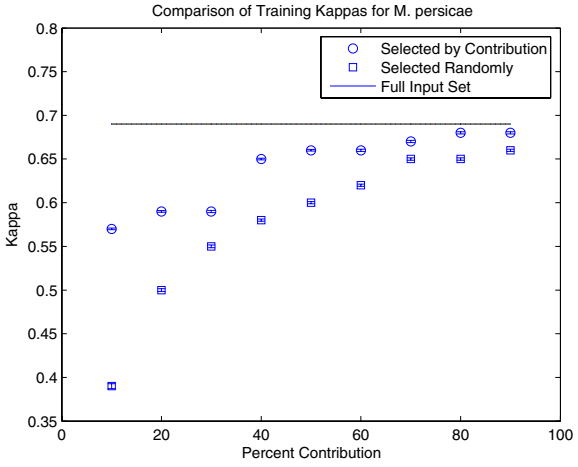

Figure 2. Mean and standard errors of training set kappas for input subsets selected by contribution and selected randomly for species *M. persicae*

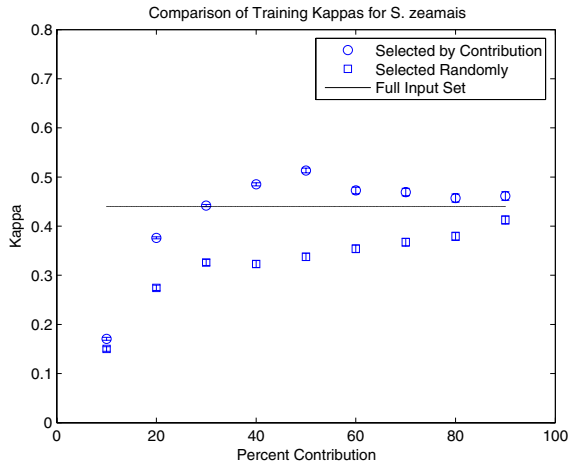

Figure 3. Mean and standard errors of training set kappas for input subsets selected by contribution and selected randomly for species *S. zeamais*

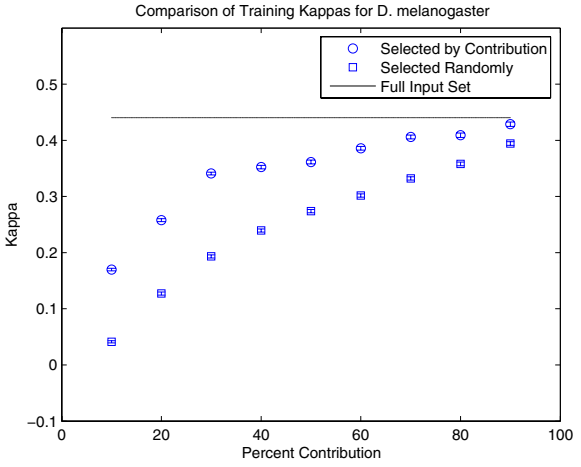

Figure 4. Mean and standard errors of training set kappas for input subsets selected by contribution and selected randomly for species *D. melanogaster*

Table 3. Similarities between variable sets.

	B. brassicae			*M. persicae*		
%	Exp.	Actual	H_0	Exp.	Actual	H_0
10	3.7	3.02/7.4	a	3.0	2.9/8.4	a
20	7.4	7.4/7.9	a	6.7	6.5/8.0	a
30	12.6	12.8/7.4	a	11.9	12.0/7.6	a
40	18.5	18.6/7.3	a	17.8	18.0/7.2	a
50	25.9	25.9/6.3	a	24.4	24.4/6.4	a
60	33.3	33.8/5.7	a	31.9	32.0/5.8	a
70	42.2	42.5/5.2	a	40.0	40.0/5.1	a
80	53.3	53.4/4.1	a	50.4	50.3/4.3	a
90	68.1	68.1/2.8	a	63.7	63.6/3.1	a
	S. zeamais			*D. melanogaster*		
%	Exp.	Actual	H_0	Exp.	Actual	H_0
10	3.7	3.9/8.6	a	3.0	2.8/8.2	a
20	7.4	7.5/7.9	a	6.7	6.6/8.1	a
30	12.6	12.7/7.2	a	11.9	12.1/7.8	a
40	18.5	18.2/7.0	a	17.0	17.2/7.4	a
50	25.2	25.5/6.3	a	23.0	23.1/6.6	a
60	31.9	31.8/5.8	a	30.4	30.4/6.1	a
70	40.0	40.0/5.1	a	38.5	38.4/5.4	a
80	48.9	48.8/4.4	a	50.4	50.2/4.3	a
90	63.0	63.0/3.2	a	64.4	64.4/3.1	a

pared to the expected level of similarity (two-tailed t-test, $p = 0.001$). If the similarities do not differ significantly from the expected, then the sampling of the feature set is assumed to be unbiased.

The results of these comparisons are presented in Table 3, where the columns labelled "%" show the percentage contribution, as in Table 2, while "Exp." and "Actual" present the expected and actual (mean and standard deviation) percentage similarities between the selected and random feature sets. The columns labelled "H_0" present the results of the hypothesis tests. In these columns, an entry of "r" indicates that the null hypothesis was rejected, while "a" indicates that the null hypothesis was not rejected.

Clearly, percentage similarities did not deviate significantly from expected values. The sampling used to select the random feature sets is therefore considered to be unbiased.

4 Discussion

The results show that the method used to determine the contribution of each input neuron (and hence the contribution of each input variable) is generally sound. The variables identified as significant for all species can be considered to be significant with some confidence.

While the null model method described here was developed for a problem from ecology, it can be applied to other problems. While the method has some computational cost, the results allow for a high degree of confidence in the statistical significance of the results of contribution analysis.

5 Conclusion

A method of validating the importance of the input features of MLP via a null-model analysis is presented. This method compares the performance of MLP trained using sub-sets of variables selected on the basis of their contribution, with the performance of MLP trained using randomly selected variables. This method was applied to the verification of the contribution of abiotic factors to the establishment of invasive insect pest species. In all cases, the factors identified by the input contribution analysis were verified by the null model analysis as statistically significant variables.

Future work in the ecology application area will include a thorough examination of the variables identified as significant, from the point of view of the biology of the insect species. Experiments will also be run using an evolutionary algorithm to select input variables.

Acknowledgements

The authors wish to acknowledge the work of Muriel Gevrey and Joel Pitt who prepared the data used in this research. This study was funded by the Centre of Research Excellence, Bio-protection, at Lincoln University.

References

[1] CABI. Crop protection compendium - global module, 5th edition. ©CAB International, Wallingford, UK, 2003.

[2] G. D. Garson. Interpreting neural-network connection weights. *AI Expert*, 6(7):47–51, 1991.

[3] N. Gotelli. Null model analysis of species co-occurrence patterns. *Ecology*, 81(9):2606–2621, 2000.

[4] L. Milne. Feature selection with neural networks with contribution measures. In *Proceedings of the Australian Conference on Artificial Intelligence AI'95, Canberra*, 1995.

[5] J. Olden and D. Jackson. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, 154:135–150, 2002.

[6] M. Watts and S. Worner. Using MLP to determine abiotic factors influencing the establishment of insect pest species. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN) 2006*, pages 3506–3511, 2006.